DOCUMENT RESUME

ED 454 875 IR 058 166

AUTHOR	Reynolds, Regina Romano
TITLE	Partnerships To Mine Unexploited Sources of Metadata.
PUB DATE	2000-11-00
NOTE	22p.; In: Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web (Washington, DC, November 15-17, 2000); see IR 058 144.
AVAILABLE FROM	For full text: http://lcweb.loc.gov/catdir/bibcontrol/reynolds_paper.htm l.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Bibliographic Records; *Cataloging; Institutional Cooperation; Metadata; Subject Index Terms
IDENTIFIERS	International Standard Serial Number; *Usability

ABSTRACT

This paper discusses the metadata created for other purposes as a potential source of bibliographic data. The first section addresses collecting metadata by means of templates, including the Nordic Metadata Project's Dublin Core Metadata Template. The second section considers potential partnerships for re-purposing metadata for bibliographic use, including metadata collected by the U.S. Copyright Office, the Library of Congress Cataloging in Publication Office, and the NSDP (National Serials Data Program) ISSN (International Standard Serial Number) registration. Other potential partners for libraries in acquiring metadata are also summarized, including DOI (Digital Object Identifier), ISBN (International Standard Book Number), and the ISTC (International Standard Textual Work Code). The third section describes the NSDP Web Template study, which compared unedited data by supplied publishers using NSDP's online ISSN application form with the completed serial records in order to assess the usability of data supplied by publishers on registration templates. The fourth section lists some potential means for increasing the usability of publisher-supplied metadata. The fifth section covers the provision of subject data, focusing on the use of controlled vocabularies. (Contains 21 notes.) (MES)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 454 875

Partnerships to Mine Unexploited Sources of Metadata

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

B. Wiggins

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Regina Romano Reynolds

Head, National Serials Data Program
Library of Congress

Final version December 2000

The old order changeth, yielding place to new,
And God fulfils himself in many ways,
Lest one good custom should corrupt the world.
"The Passing of Arthur," Alfred Tennyson

"The high cost of traditional library cataloging makes it impractical in the context of such growth and less expensive alternatives are needed for many, if not all of those resources..." 1.
"Recommendation: The Library should actively encourage and participate in efforts to develop tools for automatically creating metadata. These tools should be integrated in the cataloging workflow." 2.
LC 21: A Digital Strategy for the Library of Congress

Introduction

Let me not be quoted as saying that the "good custom" of cataloging as refined and practiced for over 100 years by the Library of Congress has "corrupted the world!" Nevertheless, as many of the conclusions of the National Academy of Science report, *LC 21*, indicate, it is becoming ever more clear that the old order of cataloging must, indeed, yield place to a new one, no doubt a multi-faceted one, with many levels, schemes, players, and partners.

If the resources that are being put onto the Web each day were, instead, being published in print form and being received into the Library of Congress, there would not be enough room left in any of our three buildings for us to hold this conference! Even though these backlogs are intangible, they nonetheless continue to grow to the point where we must think of new solutions--solutions which this conference is being held to explore.

Even though I feel we need new solutions, I do not feel that cataloging is either unnecessary or obsolete. At least not in the near-to-medium term. There may come a day when information is self-indexing, when discovery mechanisms will have progressed to the point where there is no need for this traditional library function, but, as the papers prepared for this conference indicate, we are not there yet. The thoughts I offer here are for the short-to medium term, in order to transition from where we are to where we might eventually go.

We have created our library catalogs over the course of hundreds of years, and these catalogs provide access to many resources which are not yet in digital form and some which might never be digitized. As Sarah Thomas and others have already indicated, in order to provide access to the collective wisdom the library contains we have to find ways to integrate the discovery of both traditional and Web-based

IR058166

BEST COPY AVAILABLE

resources into our catalogs. One of the gatekeepers to this process is the restriction of many catalogs to AACR and MARC-based records. CORC, with its ability to translate Dublin Core elements into MARC tags, is providing an opening wedge for non-AACR-based records.

One potential solution which I and various others have been advocating for many years is to widen the concept of levels of cataloging to encompass a hierarchy of different record levels aligned with the research (and probably monetary) value of the resource. At the highest level-rare books and materials of high research value, for example--traditional cataloging by experts in subject and descriptive cataloging would still prevail. At the lowest level-Web-based resources of moderate research value which might not otherwise receive any bibliographic control--records might be produced from publisher-supplied or secondary-source metadata which has been formatted into MARC records for inclusion in library catalogs. At levels in between these two extremes, a combination of automated and cataloger-supplied data would be used. For example, metadata-based records might be augmented by authoritative name and subject headings. Or, metadata-based records might be edited by catalogers or cataloging technicians as well as being enhanced with authoritative name and subject headings.

How can libraries obtain the metadata they need to implement such a hierarchy of cataloging levels? When I thought about this question, I realized that obtaining metadata is a lot like getting money: you can inherit it, get it the "old-fashioned way" by earning, or creating it, or-marry it!-that is, find a partner with a lot of it. And, what is the modern way to find a partner? Put an ad in the personals! So, I took the liberty of writing the following ad for the Library of Congress:

```
200-year-old Library (looks 100), mature, experienced, with mill
seeks young, exciting, digitally-savvy partners with ample metad
We can make beautiful catalogs together!  Willingness to convert
```

Although this example is obviously facetious, how libraries can obtain the metadata upon which to build a hierarchy of different cataloging levels is the subject of this paper. My premise is that metadata created for other purposes--particularly metadata created in association with existing and emerging identifiers--ISSN, ISBN, DOI--or captured from registrations such as Copyright or Cataloging in Publication, are potential sources of bibliographic data, and that non-traditional solutions such as this are the only way libraries will be able gain some bibliographic control over the explosion of Web-based resources. Partnerships with agencies which collect this metadata can provide opportunities to share libraries' experience using metadata so as to make it readily adaptable for library cataloging purposes.

Collecting Metadata by Means of Templates

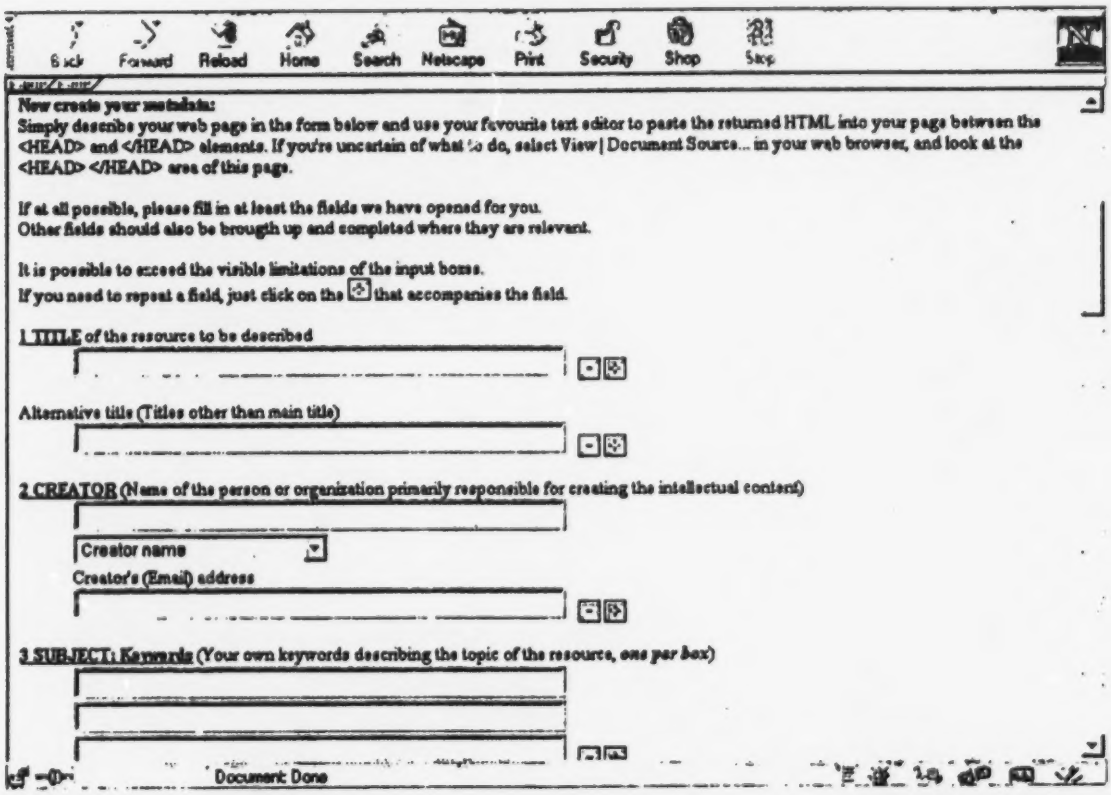
"The creation of more and better metadata--structured resource descriptions either embedded into documents themselves or external to them--is generally regarded as the best means to improve the current situation. Many specialists believe that any metadata is better than no metadata at all--we do not need to stick with the stringent quality requirements and complex formats of library catalogue systems. Instead it is possible to live with something simple, which will be easily understandable to publishers, authors, and other people involved with the publishing of electronic documents." The Nordic Metadata project. Final Report, Introduction. 3.

If all producers of Web-based resources created metadata according to a rich and accepted standard and embedded this metadata into the HTML header of their document, the "automatic" creation of metadata for library catalogs recommended by ***LC 21*** would be well on its way to becoming a reality. If XML and RDF were in widespread use and implemented by Web browsers, we would also be closer to "automatic" metadata creation. Again, we are not there yet. Simply reacting to the potential of existing technology will not yet get us where we want to go. Therefore, libraries must become more proactive in pursuit of

the means to control Web resources.

One approach in this direction has been the use of templates (online fill-in forms) to collect metadata elements and the development of programs to format the collected metadata into standardized syntax, such as Dublin Core HTML. This is the approach taken by, most notably, the Nordic Metadata Project and BIBLINK. Briefly, the Nordic Metadata project, which ran from 1996 - 1998 chose the Dublin Core element set as its metadata format, developed a template to collect and format metadata, developed metadata harvesting and indexing applications, a Dublin Core-to-MARC converter, and a URN generator. BIBLINK, a European Commission-sponsored project which ran from April 1996 - February 2000 had as its aim "to establish a relationship between national bibliographic agencies and publishers of electronic material, in order to establish authoritative bibliographic information that would benefit both sectors." 4. BIBLINK used a template as a mechanism by which publishers could send metadata to national bibliographic agencies.

Nordic Metadata Project's Dublin Core Metadata Template


Back
Forward
Reload
Home
Search
Netscape
Print
Security
Shop
Stop
New create your metadata:
Simply describe your web page in the form below and use your favourite text editor to paste the returned HTML into your page between the <HEAD> and </HEAD> elements. If you're uncertain of what to do, select View | Document Source... in your web browser, and look at the <HEAD> </HEAD> area of this page.
If at all possible, please fill in at least the fields we have opened for you.
Other fields should also be brought up and completed where they are relevant.
It is possible to exceed the visible limitations of the input boxes.
If you need to repeat a field, just click on the that accompanies the field.
1 TITLE of the resource to be described
Alternative title (Titles other than main title)
2 CREATOR (Name of the person or organization primarily responsible for creating the intellectual content)
Creator name
Creator's (Email) address
3 SUBJECT: Keywords (Your own keywords describing the topic of the resource, one per box)
Document: Done

Use of templates to collect and standardize metadata from resource creators, as in the BIBLINK project above, can result in two distinct outcomes, both highly relevant to library bibliographic control. First, by use of "crosswalks," such as those developed to convert Dublin core elements to MARC elements, MARC records can be output from data input on such templates and integrated into OPACS. Second, Dublin Core metadata coded according to HTML standards can be output and returned to the resource originator to be included in the HTML head of the resource, thus enabling cataloging tools such as OCLC's CORC and others to produce much more complete and immediately-usable MARC records.

BEST COPY AVAILABLE

Although the potential exists to extract some metadata from resources themselves, especially those where metadata has been embedded by the resource creators, one premise of this paper is that for the short-term, information elicited by carefully designed templates will be much more compatible with existing catalogs and the standards by which most of those catalogs have been created. In the report, "Project BIBLINK: Linking Publishers and National Bibliographic Agencies," Manjula Patel and Robina Clayphan state, "Records produced using the Web Interface have been of a better standard than those extracted from data already embedded in some on-line resources. The Web form has the advantage from the library's point of view, of concentrating attention on the task, limiting errors that can be made, and causing the user to create the record *following the guidelines provided.*" 5.

Potential Partnerships for Re-purposing Metadata for Bibliographic Use

Potential partners exist both within and outside of the library community. The Library of Congress is in a unique position to build on the experience of projects such as the Nordic Metadata Project and BIBLINK. The Library already receives metadata--some of it in digital form--in conjunction with at least three registration processes under its control. The U.S. Copyright Office, the Cataloging in Publication program, and the National Serials Data Program (NSDP, the U.S. ISSN center), all use registration forms on which publishers supply metadata. All three programs have mechanisms to accept at least a portion of their registrations using online forms. Outside of the Library, emerging projects with potential include the DOI for metadata created in connection with the Digital Object Identifier; registration agencies to be created in conjunction with the proposed ISO standard, the ISTC (International Standard Text Code); the ISBN Core Metadata project; and metadata collected to support OCLC's Open Name Services project. These emerging projects are very much still under development, so their potential is difficult to predict. On the other hand, opportunities for collaboration and helping to shape the metadata which is collected might be greater with emerging projects than with the well-established programs the Library of Congress is associated with.

U.S. Copyright Office

The ***LC 21*** report noted, "The Library's role in registering copyright and enforcing the mandatory deposit law creates a unique opportunity for it to collect digital information that might otherwise vanish from the historical record." 6. Along with this unique opportunity to collect digital material, the Library also has a corresponding unique opportunity to collect metadata to support not only the copyright registration but also to facilitate the bibliographic control of this material. Although at present most copyright registrations are made by applicants using paper forms, an "Electronic Registration, Recordation & Deposit System," CORDS, is under development and has been in limited use with a group of publishers who use a digitized template of the paper form to submit applications and copies of digital materials.

As CORDS is moved into a production system, as urged by the LC21 report, or as a new digital registration system is developed (another option presented by ***LC 21***), the potential for collaboration with the cataloging operation of the Library will open up in a new way, although some significant hurdles will need to be overcome. At present, cataloging of copyright registrations and cataloging for bibliographic control have been separate, parallel systems at the Library of Congress. For years, those not familiar with the details of copyright registration have wondered why these operations could not somehow build upon each other. At first glance, this seems like an obvious possibility. However, Associate Register Mary Levering explained to this author that the copyright form collects only "copyright registration facts," metadata which do not necessarily match metadata created for traditional cataloging. 7.

A preliminary comparison of the Dublin Core elements with the elements collected as part of the copyright registration process reveals the following:

D.C. Element	On Copyright Form
Title	Title
Creator	Author
Subject	-----
Description	-----
Publisher	-----
Contributor	Combined with author
Date	Date creation of work completed Date of first publication
Type	Type is denoted by choice of form, e.g., VA= visual works, PA = performing arts works, etc.
Format	Some formats require specific forms
Identifier	URL is not collected; ISSN is collected only on form SE/Group which is used only for group registration of periodicals
Source	Derivative work or compilation
Language	----
Relation	Title of collective work for works published as contributions to periodicals, serials, or collections, includes designation and pages of the periodical or serial
Coverage	----
Rights	Claimants, OPTIONALLY on some forms: name and address of person to contact re rights and permissions

Additional elements present on Copyright Registration Forms

Author's nationality or domicile
Whether the author's contribution was anonymous or pseudonymous
Author's dates of birth and death
Nature of authorship (entire text, co-authorship, compilation, translation, etc.)
Address of Copyright claimant
Previous registration
Work made for hire

The above brief comparison indicates that for five of the 15 DC elements there is no equivalent data and for various others the data is incomplete or does not map completely. Such typical bibliographic elements as subject, place of publication and publisher are totally absent. Nonetheless, because much work remains in order for the Copyright Office to have a production system for the registration of electronic materials, potential for collaboration remains. Additionally, both the Library's cataloging operations and Copyright Office have a common interest making the best use of limited staff in the face of the increasing volume of electronic materials confronting each operation. A form of collaboration heretofore seen as only desirable might soon become essential.

Associate Register Levering indicated that the Copyright Office could ask for additional elements on the registration application--especially if provision of such elements was optional-- without any change in

law or regulation. However, she also indicated concern about placing a potentially discouraging burden of information on Copyright applicants. A first step would be providing the opportunity for publishers to supply additional optional metadata which would support the Library's cataloging operation. Currently, certain copyright forms provide a box (labeled OPTIONAL) in which to provide "Name and Address of Person to Contact for Rights and Permissions." A very informal estimate by some Copyright Office staff indicates that this optional information is often provided. Levering indicated that consensus support for the inclusion of provisions for the collection of additional metadata-probably optionally-- from key organizations in the library, publishing, and information communities could lead to consideration of including these additional elements in a future digital copyright registration system. 8.

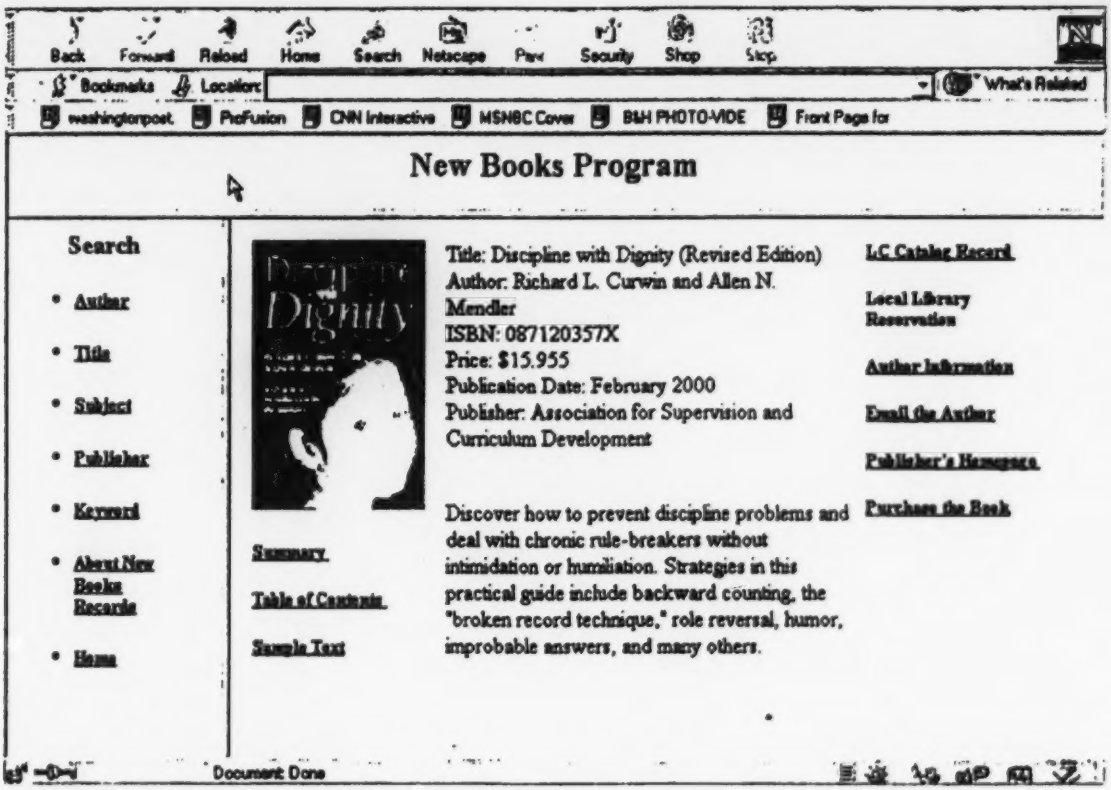
Once the minimum elements needed for a library catalog record were collected via an electronic copyright registration system template, these data could be converted into baseline records and shared with the bibliographic community as LC has been doing with other kinds of cataloging data for the over a century. The challenges to be overcome are considerable but the payoffs in terms of increased control of U.S. digital resources would be great.

Cataloging in Publication Office (CIP)

Another of the "narrow gates" through which many U.S. monographic materials pass, and thus a point for the potential capture of metadata for electronic materials, is that of CIP. The current CIP registration form elicits much more bibliographic information than the Copyright form. However, at present, Web resources are not included in the CIP program. The "Electronic CIP" program, thought of by some as pertaining to digital materials, in fact, only encompasses printed materials sent to LC in digital form but, given publishing and library control trends, that policy would seem to be destined for change in the future. Nonetheless the present "Electronic CIP" program does provide a model for the receipt of digital materials in the future since the Electronic CIP programs designed to process applications for printed books could easily be adapted for use with digital resources. Additionally, a proposal entitled "New Books" currently under development in the CIP office also demonstrates the power of the CIP program to interact with publishers for a net result with great potential for the control of electronic resources.

Under the Electronic CIP Program, publishers complete an online form and attach a digital file of the book. Using a program developed at the Library, the data on the form is converted to a draft MARC record. LC catalogers use the draft record and digital file to produce cataloging in publication records, which are later updated using donated copies of the published books. Working with galleys in digital rather than printed form has produced some distinct advantages. More CIP records are based on full text rather than simply front matter; fewer typographical errors occur because catalogers can cut and paste data from the electronic text; and quicker throughput time results from eliminating transit time in the mail and within the Library.

The proposed New Books program is still just a gleam in Cataloging in Publication Program Chief John Celli's eye--and a prototype on an internal Library server. Nonetheless, the prototype dramatically demonstrates the usability of publisher-supplied metadata to create not only basic catalog records but also to enhance such records in ways which the public is coming to expect based on their use of such Web sites such as Amazon.com.



The New Books proposal would provide for a template on the Web which participants in both the CIP Program and the Preassigned Card Number program could complete. Elements requested on the template are the traditional catalog record elements such as title, subtitle, place, publisher, date, subject, audience, ISBN, plus the potential to include a summary, reviews, table of contents, and image of the book jacket. The data input on the template and the attached files are then converted into a record that looks like a cross between a traditional catalog record and a listing on Amazon.com. Records created in this manner would, under the New Books proposal, form their own adjunct to the LC OPAC as well as reside in a space reserved for them on LC's Web page. As the materials were published and selected for inclusion in LC's collection, they would receive traditional cataloging and records would be removed from the Web site under one scenario, or simply moved to a published portion of the site under another. Although not intended, as now conceived, as a replacement for traditional cataloging, it is difficult to imagine that the potential efficiency to be realized by simply bringing this publisher-created record under name and subject control might not lead to such a future. Furthermore, although the initial New Books concept does not focus on digital materials, given the pressures on the Library of Congress brought to bear by the challenge of collecting and controlling a much greater percentage of Web-based resources in the future, it is also hard to imagine that the potential for this kind of record-creation mechanism will not be recognized and extended to digital resources. John Celli, himself, has indicated to this author that digital resources will probably have to be included if the project is to gain widespread Library support.

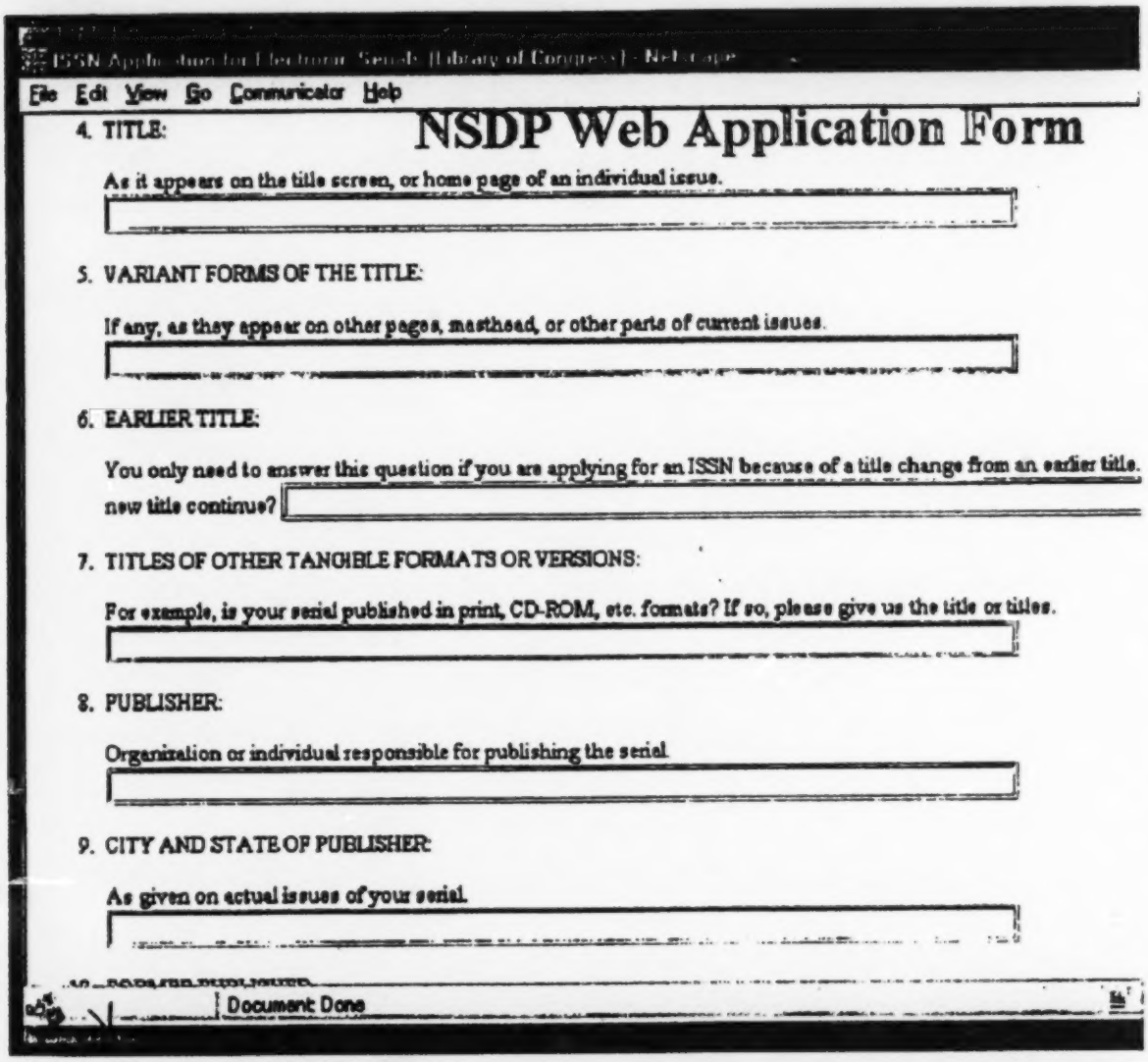
One very attractive feature of the New Books prototype is the potential for the addition of subject headings based on a pull-down menu of the "BASIC Subject Heading Codes," a list developed by the Book Industry Study Group which will be discussed at greater length in the last section of this paper. This list consists of "over 2500" subject headings--a tiny fraction of the number in the 5-volume set of Library of Congress Subject Headings (LCSH). Nonetheless, publishers' use of this list would provide controlled subject vocabulary, and either alone, or better yet, coupled with key word searching, would

BEST COPY AVAILABLE

result in some controlled vocabulary subject access without intervention by library cataloging staff.

ISSN Registrations

Since 1996 the National Serials Data Program (NSDP), the U.S. ISSN center, has used a template form on its Web site (http://lcweb.loc.gov/issn) for ISSN registration of U.S. digital serials. The online form includes the same data elements as those present on the printed application form but its use is restricted to digital serials. Use of the online form requires the publisher to provide either digital files along with the application or a URL so that ISSN center staff can view the serial in order to determine its eligibility for registration. The online form already provides many advantages to quicker and more accurate ISSN registrations and records for digital serials but it holds even more potential for the future, since both the template and conversion program are still quite basic and only the ISSN is returned to the publisher at present.


File Edit View Go Communicator Help
NSDP Web Application Form
4. TITLE:
As it appears on the title screen, or home page of an individual issue.
5. VARIANT FORMS OF THE TITLE:
If any, as they appear on other pages, masthead, or other parts of current issues.
6. EARLIER TITLE:
You only need to answer this question if you are applying for an ISSN because of a title change from an earlier title.
new title continue?
7. TITLES OF OTHER TANGIBLE FORMATS OR VERSIONS:
For example, is your serial published in print, CD-ROM, etc. formats? If so, please give us the title or titles.
8. PUBLISHER:
Organization or individual responsible for publishing the serial
9. CITY AND STATE OF PUBLISHER:
As given on actual issues of your serial.
Document Done

P functions within the CONSER program, a cooperative cataloging program for U.S. serials. As part of CONSER, NSDP creates CONSER records in the OCLC database for the serials to which it assigns

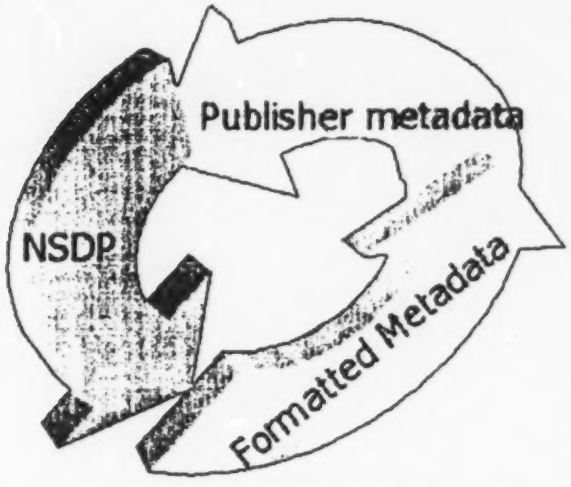
BEST COPY AVAILABLE

ISSN. The conversion program has been designed for use with OCLC. To begin with, the template program screens applicants by asking if the resource for which registration is an online serial, and whether it is published in the United States. If either of these questions is answered in the negative, an error message appears, and the form cannot be submitted. Instead, instructions are provided for how to register serials in other formats or how to contact other ISSN centers to register serials published outside of the United States. Thus, inappropriate registrations are avoided--for the most part. Of course, there is always a way for a determined applicant to subvert even the most well-designed program. For example, NSDP recently received, via the U.S. Mail, a printout of NSDP's online application form completed by a library organization in the United Kingdom. The form was accompanied by a note which indicated that the organization could not successfully submit the form online and thus were sending it by post!

When the cataloger processes the file resulting with the data supplied by the publisher, the elements on the form which contain data appropriate for a MARC record are mapped to their corresponding MARC21 fields by a program developed at LC. The cataloger then edits and augments the resulting draft record according to the requirements of the ISSN Network, AACR2 and MARC21, including the checking of authority files. The current program is quite basic--more elements could be requested from the publisher, more explicit instructions could be given to obtain better data, and better error checking could be done. Nonetheless, catalogers find the program cuts down on the keying they have to do, and, even though publishers can also mis-key data, use of the program has resulted in greater accuracy, especially in URL fields. Additionally, processing efficiencies have been realized since there is less paper to move, file, and potentially misplace; and publisher notification of the ISSN is done by an email program, thus reducing notification time and effort.

The potential for returning standardized and formatted metadata to resource creators is an exciting byproduct of the registration process. NSDP and other registration agencies are in a unique position to complete the information loop by returning metadata back to the publisher--now standardized, enhanced, and properly formatted for embedding into the head of the resource. NSDP is hoping to develop a program--or to use OCLC's CORC capabilities--to output not only a MARC21 record for cataloging but also HTML metadata for return to publishers. Publishers who are standards-aware enough to have requested ISSN might be especially good candidates for embedding standardized metadata into their resources if it is supplied to them at the time of ISSN registration. In this way, search engines, harvesters and programs such as MARC-it and CORC can return much superior results.

Linking Publishers & Libraries



NSDP receives approximately 50 ISSN applications a month using the online template. This volume of template use has taken place with no specific publicity on the part of NSDP. The number of online registrations would probably be much greater if this capability were publicized but NSDP lacks the staff to keep up with even the current number of ISSN requests. The ISSN Network is also exploring the potential of template-based self-registration by creators and publishers of digital resources. Such self-registration might be promoted in collaboration with selected publishers or information community partners as a means of the ISSN Network's strategic goal of increasing coverage of online resources beyond the limited staff capabilities of ISSN centers.

Other Potential Partnership Projects

The projects or organizations described below as potential partners for libraries in acquiring metadata which could form the basis of catalog records are all more or less in the formative stages as far as such collaborations are concerned. That many of these projects are still in the early stages presents both advantages and disadvantages. Although it is natural to want a well-established partner, it is also more difficult to bring about changes in the procedures of such partners. With partners whose operations are still under development, there will be more opportunities to influence development in ways mutually beneficial to all partners.

Before I discuss specific potential partners, I would like to mention a metadata framework--< indecs >--and a metadata standard--ONIX International--which are relevant to some of the potential partnership projects listed below. The < indecs > (interoperability of data in e-commerce systems) project "was created to address the need, in the digital environment, to put different creation identifiers and their supporting metadata into a framework where they could operate side by side, especially to support the management of intellectual property rights.... < indecs > is designed to help bridge the gap between the powerful but highly abstract technical models such as that expressed in the Resource Description

BEST COPY AVAILABLE

Framework (RDF) and the more specific data models that are explicit or implicit in sector- or identifier-based metadata schemes." 9. The indecs metadata framework has influenced the DOI Foundation's metadata model as well as the development of ONIX and other metadata projects.

The ONIX International metadata set has been produced by EDItEUR jointly with the Association of American Publishers (Washington), Book Industry Communication (London), and the Book Industry Study Group (New York). ONIX Version 1.01 states that "ONIX International is the international standard for representing and communicating book industry product information in electronic form." 10. The ONIX metadata set includes product identifiers (ISBN, EAN-13, DOI, etc.); product descriptions (author, title, edition, publisher, publishing dates, series/set, subjects); product "promoters" (annotations, prizes, reviews); and product business data (supplier restrictions, pricing).

DOI (Digital Object Identifier)

According to the DOI Foundation's Web site, "The Digital Object Identifier (DOI) is an identification system for intellectual property in the digital environment. Developed by the International DOI Foundation on behalf of the publishing industry, its goals are to provide a framework to manage intellectual content, link customers with publishers, facilitate electronic commerce, and enable automated copyright management. " 11. Although early DOI registrations did not include associated metadata, according to the DOI Handbook (5.1), "Metadata is an essential component of the DOI System, and declaration of a limited "kernel" of metadata will soon become mandatory for all DOIs that are registered." 12. Additionally, "genres," defined areas of intellectual property, will be defined to serve different communities of interest and these genres will be able to define additional metadata elements appropriate to the particular genre.

Although the DOI has not as yet become a key identifier for Internet resources, the DOI syntax was approved by the American National Standards Organization in May 2000 and published as an official standard (Z39.8402000). According to the International DOI Foundation Annual Review dated September 2000, the DOI has gained in the number of participants and collaborations during the past year. 13. If the DOI continues to grow in use and participants, it might eventually prove a valuable source of metadata for Web-based resources. A potentially key application using the DOI is CrossRef, a collaborative effort which over 50 publishers had joined as of this writing (Oct. 2000). CrossRef enables linking between citations in one journal article to the cited content in another journal, even if that content is published by a different publisher and available on a different server. Success of CrossRef could give the DOI higher visibility as a viable system.

There is potential for collaboration between the DOI Foundation and libraries. The DOI Web site's FAQs indicate an interest in collaboration with libraries. The following question and answer opens the door to such collaboration:

> Question 2.4 Are there any plans to extend the DOI to bibliographic resources, like library card catalogues, to provide a subset of bibliographic information?
>
> The current plans for the prototype are for all participants in Internet-enabled publishing to determine how the DOI will work for their purposes, and we encourage other parties to begin considering the DOI as a tool for additional functions and services, such as metadata, bibliographic data and copyright management systems. 14.

Additionally, Norman Paskin, Director of the DOI Foundation, responded to this author's question about the Foundation's interest in potential collaboration with libraries for the purpose of sharing metadata with these words, "The concept you are presenting - that 'metadata created for other purposes such as

copyright registration, ISSN registration, CIP applications, and identifier registrations could form a basis for... catalog records' -- is one which we strongly support, and which is in fact central to our efforts. Re-use of metadata is a natural consequence of current developments." 15. Paskin went on to cite several developments which would support such collaboration. First is the fact that the DOI Foundation is using principles from the indecs framework, and implementations like ONIX, as the basis of its metadata. Use of these standards would provide some common ground since LC's Network Development and MARC Standards Office is working with others on a "crosswalk" from ONIX to MARC21. Such a crosswalk would allow data supplied by publishers in connection with DOI registration to be converted into corresponding MARC21 fields. Additionally, the DOI is using the Handle System, which is the underlying system for the Copyright Office's CORDS project.

Currently the elements in DOI kernel metadata number seven: identifier, title, main creator and role, type, mode, genre. However, the DOI's support of ONIX's very rich metadata set might well result in the provision of many more elements by at least some publishers. Whether these elements would all be publicly available is a question, however, since the "kernel" metadata elements are the only elements which are definitely intended to be freely available as a lcok-up from the DOI. It will be up to each genre community to define access to other elements. However, even if such elements were not freely available to the public, libraries would be in a good bargaining position to negotiate access, perhaps in return for access to library authority files.

ISBN (International Standard Book Number)

Metadata collected in association with the registration of U.S. books for ISBN has long been published in Books in Print. Currently, ISBN agencies are in the process of exploring their future in the digital world, giving libraries the potential for another source of metadata for Web resources. In a position paper prepared by the U.S. ISBN Agency entitled, "The Digital World and the Ongoing Development of ISBN," 16. some guidelines for assigning ISBNs to digital files are listed. For example, "format/means of delivery are irrelevant in deciding whether a product needs an ISBN..." and "each format of a digital publication represents a new edition and should have a separate ISBN." There has been a recognition that some metadata elements applicable to print are not relevant for digital materials, and that digital materials may require the definition of new elements. Accordingly, the International ISBN Agency is in the process of determining core metadata elements appropriate to those materials produced using print on paper and those appropriate to digital products. This core metadata work is being done in conjunction with EDItEUR International

One problem for libraries is determining Web resources of potential value to their patrons. Since the ISBN is already in use by those publishers whose works libraries have traditionally collected, perhaps metadata collected in conjunction with ISBN registrations will provide one means of narrowing the ever-expanding universe of resources libraries have to consider for selection purposes. Another potential benefit of collaboration with ISBN is that the ISBN has both the advantages of a long-established system, as well as the advantages of being in the early stages of involvement with Web resources and thus potentially open to collaboration on metadata elements and their re-purposing as the basis for bibliographic records.

ISTC (International Standard Textual Work Code)

The International Standard Textual Work Code (ISTC) is Project 21047 under the auspices of the International Standards Organization (ISO), a step on the way to becoming an ISO standard. The purpose of the ISTC, as stated in its draft scope statement is, "to enable the efficient identification of and administration of rights to textual works, particularly in the digital information environment. The ISTC provides a means of uniquely identifying works of text in databases and other sources and for the

exchange of information about those works among authors, agents, publishers, retailers, librarians, and other interested parties on an international level." 17.

It is important for our purposes to note that the ISTC is meant to be a work-level identifier, appropriate to all manifestations of the same work. Libraries have generally cataloged different manifestations of textual works on different bibliographic records. However, with the increasing proliferation of manifestations of works, many libraries are reconsidering this practice since patrons and reference staff find the multiplication of records for the same work confusing. This may be the beginning of a movement towards describing works --at least in some cases--rather than manifestations in library catalogs. Indeed, this is already becoming the case in some libraries which are following "one record policies" by simply adding URLs for online manifestations to existing records for print manifestation. So, metadata created in conjunction with ISTC registrations may be compatible with at least some library cataloging practices.

The ISTC will require registration by publishers or other interested parties. The form of such registrations has not yet been determined but it is not far-fetched to assume some form of Web-based registration might be offered. Metadata will be collected to support such registrations. The current project draft specifies the following metadata elements: title (at least one) with appropriate title type indicated; at least one author if on record, or if not, at least one contributor to the work with their respective roles indicated; whether or not the work is derived from another work and if so, the type of derivation; in the case of a derived work, the ISTC of the source work or the title if no ISTC exists for the source work; a unique identifier for the registration of the ISTC. The developers of the ISTC recognize that there will be a need to indicate the relationship between an ISBN or ISSN and the ISTC in various applications. That section of the draft is still under development.

The ISTC will be assigned through registration agencies. It is likely that multiple agencies may be established to serve various segments of the textual works community. Such registration agencies would be potential partners for bibliographic projects in their respective areas of focus.

OCLC Open Names Project

OCLC is developing a project called "Open Name Services." The premise of the project as stated on its preliminary Web site is that "Web services should be built around names and the communities that support them." On the Web site OCLC states that it is researching "how traditional names like ISBN can be used in more Web-based services and how these names can be used to link these services." 18. The initial focus is on ISBNs but the plan is to build similar services using a variety of names, such as--potentially- ISSN, SICI, Handles, etc.

The services associated with these names will have to be supported with metadata, much of which already exists. Because OCLC's base is the library community, the potential for collaborating in the collecting and sharing of metadata associated with the project would seem to be great. One of the project's supporting statements indicates "this would allow many of the traditional library services to be provided along with many new services." In fact, OCLC is actively soliciting the participation of the library community in this project, "We need groups like OCLC members and publishers to agree to open names. We then need organizations to step forward and commit to services on these names." Thus there exists potential for libraries to work together with OCLC on possible collaborations.

NSDP Web Template Study

In order to assess the usability of data supplied by publishers on registration templates, a study was carried out comparing unedited data by supplied publishers using NSDP's online ISSN application form

with the completed CONSER serial records resulting from editing and updating by a cataloger in NSDP. The cataloger responsible for making assignments to electronic serials was asked to save "before" and "after" printouts for post-publication ISSN requests. At the time of the study, 220 records had been saved. A 25% random sample was taken, resulting in analysis of 55 records. Seven elements were chosen for comparison: Title, Variant Title, Frequency, Publisher, Place, Designation, and URL. A system was devised for scoring the data supplied by the publisher as either a "Match," "Close Match," or "No Match" when compared to data on the final cataloged record. One person did all of the scoring. "Match" constituted an exact match. "Close Match" was used for cases where there were only differences that would not affect searching or identification, such as capitalization, punctuation, or full form vs. abbreviated form. These differences were differences only in form-and minor ones at that-with no difference in fact. In the table below, "Match" and "Close Match" were added together to produce a combined score.

During the course of the scoring it became clear that the element, "Variant Title," presented scoring difficulties because sometimes publishers supplied variants that the cataloger did not include at all, while other times the cataloger did include the variant but constructed the variant in a different form. Because the scoring pattern for this element would not match that of the other elements, the element, Variant Title, was dropped.

The element with the highest percentage of exact matches was Frequency with 73% matches, and 16% close matches, for a combined score of 89%. In the case of Frequency, capitalization of the first letter of the frequency designation was ignored, and frequencies that differed only in capitalization were scored as matches. Although in some cases the difference between the data supplied by the publisher and that on the finished catalog record might be considered subtle by some, e.g., 6 times a year vs. bimonthly, this kind of difference was scored as "No Match" because in cataloging terms and in some serials check-in systems these are regarded as two different frequencies.

The element with the next highest number of matches was URL with 65% Matches and 24% Close Matches for a combined score of 89%. The Close Matches were mostly cases where the publisher supplied a tilde or spacing underscore and the cataloger had to convert those characters into their hex equivalents to be acceptable in the OCLC system. A new version of the NSDP conversion program now performs that conversion so today the percentage of exact matches would be 89%. The "No Match" cases occurred when the URL provided by the publisher included one or more typos, such as the use of capital I for the numeral 1; where the URL was not provided; or where the cataloger entered a URL specific to the serial while the publisher supplied a URL for the entire Web site on which the serial appeared.

The next highest combined score--82%--was for the Title element, surprisingly so, since serial catalogers have the perception that what the publisher considers to be the title often differs from what the cataloger considers to be the title. Although the combined score was relatively high, the "Match" score, 13%, was the lowest of any element because for this element only, capitalization was taken into account when determining "Matches." Capitalization of titles in catalog records does not follow standard grammatical rules but is nonetheless considered important for catalog record consistency and interpretation. Catalogers feel they must correct the capitalization supplied by publishers. Thus, for this element there were 69% "Close Matches." The 18% of the cases which fell into the "No Match" category consisted of cases where the publisher included what the cataloger considered to be a subtitle in the title field, or vice versa. The worst match, interestingly enough, was found on an application form for a serial with a generic title. The application form was completed by a monograph cataloger.

Place of publication resulted in 29% "Matches" and 51% "Close Matches" for a combined score of 80%. The Close Matches were usually the result of the publisher's inclusion of a full form of the place name which the cataloger abbreviated, or vice versa. Also, sometimes the place information supplied by the

BEST COPY AVAILABLE

publisher varied in fullness from what resulted after editing by the cataloger. However, in 20% of the cases there was no match. Sometimes, no place was supplied on the form, in other cases the place supplied by the publisher was entirely different from the place the cataloger used, and in a few cases multiple places were given by the publisher while the cataloger chose only one.

Designation, the numbering or dating scheme used by the publisher to identify individual issues, had the next highest combined score: 75% , with 30% Matches and 45% Close Matches. Close Matches varied from what was supplied by the cataloger in the use of abbreviations and in whether enumeration or chronology or both were chosen. In one of the No Match cases, the publisher had supplied Vol. 1, no.1 as the designation of the first issue, whereas the cataloger edited the publisher's statement to read Vol. 1, no. 1-2. In other cases the caption was different, e.g., "issue" vs. "number." Cataloging rules require the designation to be transcribed as it appears on the publication.

Ironically the "publisher" element was the element with the lowest combined score: 44%, comprised of 40% Matches and 9% Close Matches. This high percentage of discrepancies resulted from different interpretations of the meaning of "publisher," especially for digital works. In several cases a personal publisher was given instead of a responsible corporate body. In other cases, multiple bodies were given when only one was chosen by the cataloger. In one case, a Web design company was given instead of the corporate body responsible for the content.

Ways to Obtain More Usable Data

Although information obtained from the publisher in the above study was often factually correct, and would have been acceptable in a typical database, the highly prescriptive cataloging rules in AACR2 as interpreted by the Library of Congress require even factually correct data to be edited by a cataloger for conformity to the rules. In many cases, the rules prescribe exact transcription of the data as it is presented on the publication. In cases where the transcription does not have to be exact, the ways in which what appears on the catalog record can differ from what appears on the publication (e.g., by being abbreviated, by omission of portions of the data) are also prescribed. Although these highly prescriptive rules seem to argue against the creation of even baseline catalog records by any automated means, the results of the above study would seem to indicate that records based on metadata supplied by publishers still show potential. There are various means by which data requiring less editing for general conformity to cataloging rules could be obtained. And, of course, there is also the potential for creating certain categories of records such as the "metadata records" which are the topic of this paper, which will be acknowledged as not following cataloging rules.

Following are some potential means for increasing the usability of publisher-supplied metadata:

1. Provide pull-down menus: for elements like Frequency where a list of MARC values exists, pull-down menus could be used to limit choices to those compatible with cataloging practice. Pull-down menus could also enable publishers to supply data which could be converted into MARC21 fixed field codes (country codes, codes to indicate current or ceased, start and end dates, etc.). NSDP catalogers now must add these codes to each record.
2. Include more specific instructions: for frequently misunderstood elements like "Publisher," clearer explanations can be given for how to supply the element. Examples taken from common situations could be given. Instructions could be provided for what to supply in cases of multiple publishers, or in the case of a personal and corporate publisher. Instructions could be given for how to treat Web designers.
3. Develop better conversion programs: just as the problem with tildes and underscores was solved by having the conversion program supply the correct hex values, programs could be designed to correct capitalization in titles or frequencies even if editing by the cataloger would still be required in some

BEST COPY AVAILABLE

cases, particularly with capitalization in titles.

4. Provide interactive instructions: programs could be developed to "talk" a user through the application process, allowing him to ask for help, and giving feedback when certain responses were made.
5. Change certain cataloging rules: rules could be examined for practices out of sync with common practice, e.g., capitalization of words in titles, standardization of when to use abbreviations and which abbreviations to use regardless of how the word appeared on the publication.
6. Change descriptive cataloging practice: abandon description rooted in paragraphs on a catalog card where information not presented in certain places on resources is relegated to notes, in favor of a database approach where all information is of equal importance and values are supplied for a list of elements, akin to Dublin Core.

Preben Hansen of the Nordic Metadata Project is in the process of analyzing responses to a survey that has been available on the Web for almost three years soliciting input regarding use of the online Nordic Metadata template. Although specific results of the survey analysis are not yet available, Hansen indicated that the reactions to the template have been mostly positive. Comparison of Hansen's survey results with those obtained in the NSDP study is planned when those results become available.

Provision of Subject Data

The NISO draft standard Z39.85, Dublin Core Metadata Element Set, includes the following comment regarding the subject element, "Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme." 19. The unwieldy and even inappropriate results generated by most Internet search engines amply demonstrate the virtues of using controlled vocabularies for subject metadata.

However, subject analysis is one of the most expensive parts of the cataloging process and the one requiring the highest level of staff to perform it. In order to facilitate the provision of subject data using controlled vocabulary, the Nordic Metadata Project included in conjunction with its template, links to "all (as far as we know) vocabularies, e.g., controlled keyword lists, thesauri, classification systems, authority lists, general vocabulary system et cetera, which are freely available for navigation on the Web." 20. The template Web site includes a list of links to over 100 general and specialized tools, a list which would certainly frighten all but the most intrepid publishers. Even access only to the Library of Congress subject and name authority files were provided (these files became inaccessible-at least for the time being-when the Library's ILS went online in August 1999) publishers would still be faced with an enormous number of choices (5 large volumes worth) and a complex system of heading construction which it takes trained catalogers years to learn. For some time I have been advocating the development of a subset of LCSH which could be used as a tool by which publishers could supply subject metadata which would be fully compatible with LCSH. This would be a considerable undertaking but one with great potential benefit.

Another approach might be to use a subject list already in use by the publishing community. Such a list is the Book Industry Study Group's "BASIC Subject Heading Codes," a list which includes "over 2500 codes that may be assigned to books for bibliographic classification purposes, shelving in retail stores and searching online databases." 21. CIP's prototype New Books template includes access to the BASIC codes. Although the BASIC codes and the subjects they represent are more general than fully-subdivided LC subject headings, LCSH equivalents could most likely be determined for most of the BASIC codes. In this way, publishers could provide subject information using terms familiar to their industry while libraries could incorporate publisher-provided subject data into their catalogs using terms from their own authority files. An analysis of the potential of the BASIC codes to be converted into equivalent LCSH terms would be a useful study for a Library of Congress intern or research scholar to perform.

BEST COPY AVAILABLE

Finally, although this might seem to be testing the limits of what even the most motivated publishers might be inclined to provide, access to Library of Congress name authority files could also be made available to online template users-complete with basic instructions--so that, when possible, names of creators and contributors could be provided in authoritative form.

Conclusion: Call to Action

Libraries can no longer afford the luxury of acting as if they are the only organizations capable of describing resources. Catalog users, as studies have shown do not find significance in, or even understand, much of the information that catalogers labor to provide in a highly prescribed manner. The concept of exact transcription from a publication is particularly problematic for online resources which can change their appearance from one viewing to the next! Through the means described here and through other means suggested at this conference and elsewhere, libraries can and must make use of metadata created for other purposes to help bring some measure of bibliographic control to the ever-growing numbers of digital resources of interest to library patrons. Librarians need to share our expertise and help shape the development of metadata standards and metadata collection templates. Librarians need to share our name and subject heading expertise and authority files. Librarians need to collaborate to solve a common problem for the benefit of all in the publishing, library, and information communities. Librarians need to collaborate, not replicate. Librarians need to be partners, not competitors. There are more than enough resources to go around!

The potential partnerships described here are not intended to be exhaustive or prescriptive. Rather, they are meant to be illustrative of potential, some greater, some perhaps less. These examples are intended to provoke discussion and dialogue, which it is hoped will result in proposals that have the potential to result in control over a larger proportion of Web resources than might otherwise be possible using traditional means and resources.

What is offered here can be reduced to some take-away principles rather than a strict blueprint:

- A hierarchy of catalog record levels-with the lower levels based on publisher-supplied metadata--may be needed to bring Web-based resources under bibliographic control .
- Metadata created for other purposes can be re-purposed for library use.
- Resource creators and producers can create usable metadata, especially with librarians' help in developing standards, templates, guidelines, instructions-and motivation!
- Libraries cannot take sole responsibility for description of Web resources, but we can help, lead, guide and share expertise.

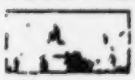
Every day thousands--if not tens of thousands--of new Web resources appear. Every day our invisible cataloging backlogs grow. The time to begin the "new order" is now!

NOTES

1. Committee on Information Technology Strategy for the Library of Congress, LC 21: A Digital Strategy for the Library of Congress. Prepublication copy. Washington, D.C.: National Academy Press, July 26, 2000, 5-6. http://www.nap.edu/books/0309071445.html
2. LC 21, 5-15.
3. Juha Hakala et al., Nordic Metadata Project. Final Report, Helsinki University Library, July 1998, 1. Introduction, http://www.lib.helsinki.fi/meta/nmfinal.htm

4. BIBLINK Project home page, http://hosted.ukoln.ac.uk/biblink/
5. Manjula Patel and Robina Clayphan, "Project BIBLINK: Linking Publishers and National Bibliographic Agencies." Proceedings: Internet Librarian 2000, March 2000 http://www.ukoln.ac.uk/metadata/publications/biblink/proj-biblink.html
6. LC 21, 3-7
7. Mary Levering, Associate Register of Copyrights, personal communication, Sept. 2000.
8. Levering, personal communication, Sept. 2000.
9. INDECS Framework, http://www.indecs.org
10. ONIX International, Version 1.01. http://www.editeur.org/onixfiles.html
11. DOI home page, http://www.doi.org
12. DOI handbook, 5.1. http://www.doi.org/handbook_2000/index.html
13. International DOI Foundation Annual Review, The Foundation, Sept. 2000. http://www.doi.org/publications.html
14. DOI home page, FAQ. a href="http://www.doi.org/publications.html
15. Norman Paskin, personal email October, 2000.
16. U.S. ISBN Agency, "The Digital World and the Ongoing Development of ISBN," New Providence, N.J., R.R. Bowker Company. http://www.bowker.com/standards/home/isbn/digitalworld.html
17. ISOTC 46/SC9 Working Group 3 for Project 21047: International Standard Textual Work Code, p. 1 http://www.nlc-bnc.ca/iso/tc46sc9/istc.htm
18. Open Name Services home page, http://names.oclc.org
19. NISO Draft Standard Z39.85-200X, Dublin Core Metadata Element Set (DCMES), http://www.niso.org/Z3985.html
20. Nordic Metadata Project. Final Report, p. 16.
21. BASIC Subject Heading Codes. http://www.bisg.org/subjectheadorder.html

---



Library of Congress
Comments: lcweb@loc.gov (December 27, 2000)

BEST COPY AVAILABLE

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate

Regina R. Reynolds

Head, National Serials Data Program
Library of Congress
101 Independence Ave., S.E.
Washington, DC 20540

Partnerships to Mine Unexploited Sources of Metadata

About the presenter:

Regina Romano Reynolds has been head of the National Serials Data Program, the U.S. ISSN center, since 1992. Reynolds has worked at the Library of Congress since 1976 and has spent much of her professional career explaining and promoting ISSN use to publishers and the information community. Reynolds has ah M.L.S. (Beta Phi Mu) from the University of Michigan. She is active in the American Library Association and the North American Serials Interest Group where she is a frequent author and presenter on topics in serials, standards, and electronic resources. Reynolds was the 1999 recipient of the Bowker/Ulrich's Serials Librarian Award. She is also actively involved in the revision of AACR2 to accommodate seriality and electronic resources as well as in the international harmonization of cataloging rules and standards.

Full text of paper is available

Summary:

If the library catalog is to play any role as a portal to Web resources, new means have to be developed to bring the ever-increasing number of Web resources of interest to library patrons under some kind of bibliographic control. Traditional cataloging of published textual materials, which has

BEST COPY AVAILABLE

Cataloging Directorate Home Page

Library of Congress Home Page

been largely monolithic to date, will have to be progressively subdivided into a hierarchy of different record levels aligned with the research (and probably monetary) value of the resource. At the highest level, traditional cataloging will still prevail. At the lowest level, records might be produced from publisher-supplied or secondary-source metadata which has been formatted into MARC records for inclusion in library catalogs. These metadata-based records could also be selectively edited by trained catalogers and optionally enhanced with authoritative name and subject headings. OCLC's CORC program is one opening wedge to the entry of such non-AACR-based records into shared databases and library catalogs.

To realize fully the potential of such metadata-based catalog records, new partnerships and new sources of cataloging data have to be explored and exploited. Metadata created in association with existing identifiers such as the ISBN and ISSN, and metadata planned to support emerging identifiers such as the Digital Object Identifier (DOI) and the developing identifier, the ISTC (International Standard Text Code), are potential sources of bibliographic data which libraries can convert, or convert and enhance to produce MARC records. Non-identifier-based publisher registration procedures such as CIP, Copyright, and others might also yield useful data. As all of these registration procedures are increasingly completed electronically, they yield data which are highly manipulable, enhanceable and convertible.

In addition to exploring sources of metadata, especially metadata supplied by publishers as part of registration procedures, this paper will examine ways in which such registration procedures could be modified to better provide libraries with needed data. Such modifications include addition of elements needed for library cataloging, and provision of instructions which will result in publishers providing data in more standardized ways. With the increasing use of online forms, interactive programs could be developed to "talk" publishers through the process of completing registration forms in such a way as to make them more useable for conversion to basic catalog records. Finally, ways for publishers to provide subject information will be explored.

The potential for creation of catalog records based on publisher-supplied metadata will be illustrated using data from a study of records created by the National Serials Data Program (NSDP). NSDP, the U.S. ISSN Center, uses an online form for ISSN registration. Publishers complete the form according to instructions supplied by NSDP. Data from the online form is converted to a draft catalog record which is then edited and enhanced by professional catalogers. Results of a study of the usability of information supplied by publishers on the ISSN application form, and the editing required on NSDP records produced by conversion from the online application form, will be presented.

Library of Congress
May 9, 2000
Comments: lcweb@loc.gov